On Statistical Multiplexing, Traffic Mixes, and VP Management *

C.–F. SU AND G. DE VECIANA
Department of Electrical and Computer Engineering
University of Texas at Austin
E-mail: {fong,gustavo}@globe.ece.utexas.edu

Abstract

ATM-based integrated services networks are likely to rely on the Virtual Path (VP) concept as an intermediate resource management layer wherein key decisions concerning resource allocation, sharing, and flow aggregation are made. In this paper we consider the impact that statistically multiplexing heterogeneous services on VP connections will have on network design and management. Based on simple models we consider several questions with perhaps surprising answers including the following. Given two traffic types with different quality of service requirements, should one segregate such flows on their own VPs, or is it to the network's advantage to multiplex the flows on a single VP guaranteeing the most stringent QoS requirement? Assuming two VPs have been set up between a given origin-destination pair and heterogeneous flows are to be carried, how should one route the connections to achieve good performance?

1 Introduction

Asynchronous Transfer Mode (ATM) has been designed to meet the possible needs of integrated broadband communication networks. This technology is based on multiplexing and switching cells transported on virtual channel connections (VCCs). Virtual path connections (VPCs) allow for joint handling of a bundled VCCs and can serve as an effective way of reducing complex switching in a core network. The VP layer is in fact likely to serve as an intermediate resource management layer, wherein key resource allocation decisions are made on a somewhat slower time scale than typical connection times. Indeed, one can use the VP layer to simplify call admission control, routing, and to segregate traffic based on QoS, traffic characteristics, or service classes. This paper addresses the question of whether or not segregating heterogeneous traffic with different QoS requirements on separate VPs is desirable. We shall see that traffic heterogeneity plays a critical role in multiplexing, and careful allocation of traffic mixes is essential in achieving good performance.

Similar care is needed in making routing decisions in a heterogeneous environment. There is much research and experience with routing policies in circuit-switched networks but one might ask if these principles will extend to multiservice networks. In circuit-switched networks, there exists a clear separation among connections, and the reserved *resources* for each connection are well defined from the start. Schemes such as selecting the "least-loaded path" can be useful for routing in single service networks because they tend to balance the traffic load across the network and minimize the blocking probability [4]. However, in a multiservice network, it has been suggested that "worst-loaded path" might be preferable in order to leave room on the "least-loaded path" for the connections with high bandwidth requirements [12]. Various other routing principles such as trunk reservation have been investigated, and these are likely to also play a role in integrated services networks, see e.g., [7].

In this paper we consider the impact that heterogeneity and statistical multiplexing might have on the design and performance of simple routing decisions, such as selecting among two possible VPCs that have already been dimensioned. Although we assume capacity has been partitioned among VPCs one would nevertheless hope that a moderate degree of statistical multiplexing can be achieved by sharing the resources allocated to VPCs. One difficulty in considering the role of statistical multiplexing in such systems is that the "effective bandwidth" required for each traffic stream may be dependent on the current load and capacity of the system, see e.g., [12]. A simple example can illustrate this and show how it might in turn impact routing policies. Consider a link shared by two types of traffic whose cell arrival rates are for simplicity modeled by Gaussian random vari-

*This work is supported by a National Science Foundation RIA Grant NCR-9409722, Career Grant NCR-9624230, and Southwestern Bell Corp./ TRC.

0-7803-4386-7/98/$10.00 (c) 1998 IEEE

ables with means m_1,m_2 and variances σ_1^2,σ_2^2 respectively. Suppose the link currently has n_1 and n_2 ongoing connections of each type. It can be shown [9] that the capacity requirement is then roughly given by

$$c(n_1,n_2) = (n_1 m_1 + n_2 m_2) + k\sqrt{(n_1\sigma_1^2 + n_2\sigma_2^2)},$$

where k is an overall QoS parameter related to a link overflow probability. The bandwidth required for an additional connection of Type 1 can be approximated by[1]

$$\frac{\partial c}{\partial n_1} \approx m_1 + \frac{1}{2}\, k\, \sigma_1^2\, (n_1\sigma_1^2 + n_2\sigma_2^2)^{-\frac{1}{2}}. \qquad (1)$$

Note that the key factor determining the marginal bandwidth requirement is the variance of the aggregate traffic $n_1\sigma_1^2 + n_2\sigma_2^2$ currently on the link, suggesting that it may be of interest to monitor the variance in order to estimate bandwidth requirements for incoming sessions. Based on this example we conclude that, it may be "cheaper" (consuming less additional bandwidth) to route a new connection through a link whose current aggregate variance is large. In turn by selecting routes with minimum marginal bandwidth requirements one might make more resources available to incoming connections or other types of services.

This argument is in sharp contrast to typical routing policies that try to balance the loads on the network. Indeed a naive interpretation of (1) suggests that we might want to generate imbalances on various routes because they may result in better multiplexing, and are thus more "economical." In the usual circuit-switched environment, the bandwidth requirement for each traffic stream is a constant independent of other traffic currently sharing the links. By contrast, in packet-switched networks statistical multiplexing and the traffic mixes on the link will affect the bandwidth requirements and thus judicious routing of connections may improve the system performance. In this paper we shall show that this is indeed the case. Routing decisions which account for the impact of statistical multiplexing, and the relative loads of various traffic types can significantly improve performance.

The balance of this paper is organized as follows: §2 discusses the role of statistical multiplexing and nonlinear call admission regions of heterogeneous traffic. A simplified VP layout problem is analyzed in §3. Routing issues are discussed in §4 and followed by conclusions.

2 Statistical Multiplexing, Traffic Mixes, and Admissible Region

We first review bufferless statistical multiplexing based on well known effective bandwidth results, see e.g., [8]. Suppose N i.i.d. traffic streams are carried on a bufferless link, and each stream has an arrival rate $A_i(t), i \in [1,...N]$. Assume that the link capacity is c and the overflow probability requirement is δ. Based on Chernoff's bound [1], the overflow probability is upper-bounded by

$$\mathbb{P}(\sum_{i=1}^{N} A_i(t) > c) \leq \exp(-\sup_{\theta>0}[c\theta - N\Lambda(\theta)]) \leq \delta, \qquad (2)$$

where $\Lambda(\theta) = \log(\mathbb{E}[\exp(\theta A_i(t))])$ is the logarithm of the moment generating function of $A_i(t)$.

In order to meet the overflow probability requirement, the link capacity c needs to satisfy the following inequality:

$$\begin{aligned} \sup_{\theta>0}(c\theta - N\Lambda(\theta)) &\geq -\log(\delta), \\ \Rightarrow \Lambda^*(\frac{c}{N}) = \sup_{\theta>0}(\frac{c}{N}\theta - \Lambda(\theta)) &\geq -\frac{\log(\delta)}{N}. \end{aligned} \qquad (3)$$

The minimum capacity $\frac{c}{N} = \alpha(\delta)$ which satisfies (3) is the "effective bandwidth" of each traffic stream subject to the QoS requirement δ on the bufferless link. Moreover, $\Lambda^*(x)$ is an increasing function on $[m,+\infty]$ [10], where $m = \mathbb{E}[A_i(t)]$ and $\Lambda^*(m) = 0$, see Fig. 1. When the number of connections becomes large, the right hand side of (3), $-\frac{\log(\delta)}{N}$, decreases and $\alpha(\delta)$ moves toward m, showing that the "economies of scale" will eventually allow the resource allocation to be based on source's mean rate.

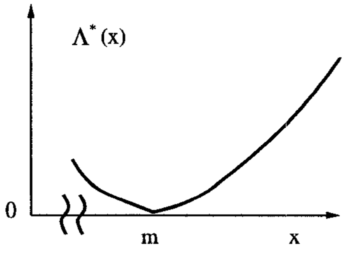

Figure 1: A typical $\Lambda^*(c)$.

One can generalize the results to multiple traffic types and analyze the statistical multiplexing among heterogeneous traffic. Let us define the *admissible region* $\mathcal{A}(c)$ for a bufferless link with capacity c as the set of all combination of connections $(n_1,...,n_J)$ from J traffic types which can be carried simultaneously subject to an overflow probability constraint δ. For the sake of simplicity and illustration, we

[1] This approximation is based on the derivative of $c(n_1,n_2)$ with respect to a continuous variable n_1. It can be shown to be accurate when the aggregate variance is high.

consider the case of two traffic types in the following discussion.

Based on Chernoff's bound in (2), a pair $(n_1, n_2) \in \mathcal{A}(c)$ should satisfy the following inequality:

$$\sup_{\theta>0}(c\theta - n_1\Lambda_1(\theta) - n_2\Lambda_2(\theta)) \geq -\log(\delta),$$

where c is the link capacity and $\Lambda_1(\delta), \Lambda_2(\delta)$ are the logarithms of the moment generating functions of Type 1 and Type 2 traffic respectively. It has been shown in [6] that the complement of the admissible set is a convex region, and it was suggested in [2, 8, 5] that a linear approximation could be used to represent the boundary of admissible region. However, this linear approximation of admissible region boundary is not always accurate.

In Fig. 2 we plot the admissible region of a bufferless link with capacity 25 and its linear approximations. The two traffic types are on/off which have peak rates 1 and 0.5 with mean-to-peak ratios 10% and 90% respectively. Fig. 2 shows that the admissible region's boundary is convex and that linear approximations need not be accurate, suggesting that the effective bandwidth of each traffic type is indeed state-dependent and sensitive to traffic mixes. For comparison, we increase the link capacity 4 times, and the new admissible region is plotted in Fig. 3. As can be seen in the figure, the admissible region is more "linear" and the linear approximations are likely to be more accurate due to better multiplexing among different traffic types.

Such nonlinearities have been recognized, see e.g., [11], where they are called the *diversity cost* and an attempt was made to quantify them. It was pointed out in [3] that the total allocated bandwidth for a single type of aggregate traffic needs to exceed a critical value in order to make the traffic "statistically-multiplexable." Thus a minimum capacity is required to see the "economies of scale." The results in [3] also showed that multiplexing "statistically-multiplexable" and "nonstatistically-multiplexable" will create a nonlinear admissible region which can not be effectively approximated by a linear hyper-plane.

Our premise herein is that the traffic mixes impact multiplexing, even though the impact may be neglected when the number of connections (or system capacity) becomes large. In carrying multiple traffic types on segregated VPs, where the bandwidth and number of connections are both moderate, the admissible region is indeed nonlinear due to the diversity in burstiness and insufficient statistical multiplexing. The traffic composition will thus be an important factor in bandwidth dimensioning and management of VPs.

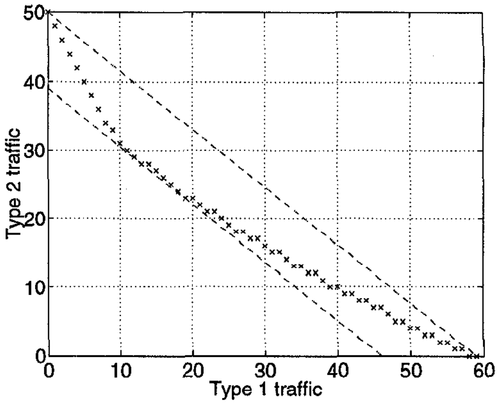

Figure 2: An admissible region and its linear approximations.

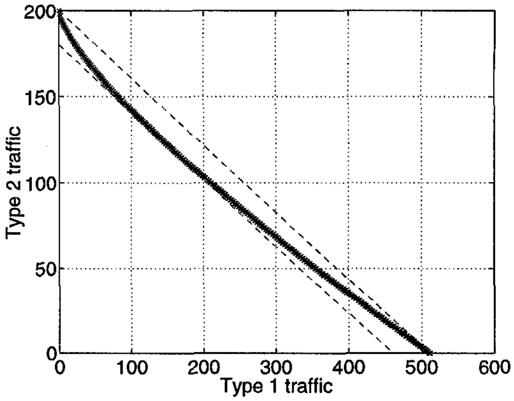

Figure 3: Another admissible region.

3 Integration or Segregation of Traffic?

It is generally believed that one should segregate traffic with different QoS requirements on their own VPs. The intuition is that if we multiplex traffic with different QoS requirements on the same VP, then the overall QoS for the VP shall be the most stringent QoS requirement. By providing a QoS that is more stringent than necessary to some traffic streams, we waste network resources. However, due to the nature of statistical multiplexing, it may still be more "economical" to put all traffic on the same VP. In the following we use an example of two traffic types to illustrate the roles of statistical multiplexing and traffic mix.

Suppose there are N total connections which consist of a fraction f_1 of Type 1 flows and a fraction f_2 of Type 2 flows, where $f_1 + f_2 = 1$.[2] In order to get a qualitative understanding, we shall resort to Gaussian traffic models for which an explicit expression for bandwidth requirements exists. The

[2] For simplicity we assume f_1, f_2 are real numbers even though they should be restricted to multiples of $\frac{1}{N}$ such that Nf_1, Nf_2 are integers.

cell arrival rates of each traffic type are modeled by Gaussian random variables with mean and variance (m_1, σ_1^2), (m_2, σ_2^2) respectively. We assume that the two traffic types are carried by a bufferless link and require cell loss ratios of 10^{-6} and 10^{-3} respectively. The goal is to decide whether to partition a link into two segregated VPs, or form a single shared VP, and we need to assess the bandwidth requirements for the two options.

The total required capacities for these two cases are:

$$c_1 = N(f_1 m_1 + f_2 m_2) + k_1\sqrt{N}\sqrt{f_1\sigma_1^2 + f_2\sigma_2^2} \qquad (4)$$

$$c_2 = N(f_1 m_1 + f_2 m_2) + \sqrt{N}\left[k_1\sqrt{f_1\sigma_1^2} + k_2\sqrt{f_2\sigma_2^2}\right] \qquad (5)$$

where k_1 and k_2 are the QoS parameters. For the Gaussian model, the tail distribution can be captured by the deviations from mean, and k_1, k_2 correspond to the standard deviation multiples [9]. The bandwidth requirements of a single shared VP and segregated VPs are shown in (4) and (5) respectively. Without loss of generality, we assume $k_1 > k_2$. For the aforementioned QoS, k_1 and k_2 are 4.7534 and 3.0902 respectively. We are interested in a condition making $c_1 \leq c_2$, so it is advantageous to form a single VP and give Type 2 traffic a better QoS, rather than setting up two VPs. In other words, the benefit of statistical multiplexing outweighs the loss in over-provisioning for a better QoS.

Surprisingly, the condition depends only on $\frac{\sigma_1}{\sigma_2}$ and f_1, where N plays no role. Indeed for $c_1 \leq c_2$, we need that

$$k_1\sqrt{f_1(\frac{\sigma_1}{\sigma_2})^2} + k_2\sqrt{1-f_1} \geq k_1\sqrt{f_1(\frac{\sigma_1}{\sigma_2})^2 + 1 - f_1},$$

which can be rewritten as

$$\frac{\sigma_1}{\sigma_2} \geq \frac{k_1^2 - k_2^2}{2k_1k_2}\sqrt{\frac{1-f_1}{f_1}}. \qquad (6)$$

Hence to decide whether or not to integrate two types of traffic on the same VP, one needs to assess the ratio of their variance, versus a function of their QoS requirements and the traffic mix.

In Fig. 4 we plot the threshold of $\frac{\sigma_1}{\sigma_2}$ as a function of the traffic mix f_1 with the aforementioned k_1 and k_2. The threshold defines the *integration* and *segregation* regions. For example, for $\frac{\sigma_1}{\sigma_2} = 0.6$, we should form a single VP when the fraction of Type 1 traffic exceeds 0.35. Otherwise, it is more efficient to have two VPs with different QoS.

Based on the regions shown in Fig. 4 it is clear that when f_1 is small, the ratio of $\frac{\sigma_1}{\sigma_2}$ needs to be large in order to make integration beneficial. An interpretation for this might be that we waste a larger amount of bandwidth in bringing a better QoS to Type 2 traffic (i.e., integration) when f_1 is small. Thus only when Type 1 traffic is "bursty", i.e., has high variance, can the benefit of better multiplexing outweigh the waste of bandwidth due to QoS over-provisioning. Therefore, the threshold $\frac{\sigma_1}{\sigma_2}$ should be larger when f_1 is small, as shown in Fig. 4. By contrast when f_1 is large, the Type 2 traffic becomes less significant, so the threshold $\frac{\sigma_1}{\sigma_2}$ becomes less stringent, i.e., *integration* is desirable. The trade-offs of integration are captured by the curve in Fig. 4.

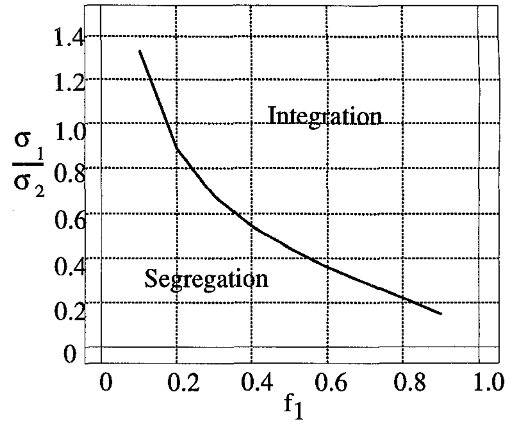

Figure 4: Traffic mix vs. $\frac{\sigma_1}{\sigma_2}$.

Note that $\frac{k_1^2-k_2^2}{2k_1k_2}$ serves as a scaling factor for the the curve in Fig. 4. If k_1 and k_2 are close, then the threshold $\frac{\sigma_1}{\sigma_2}$ for integration becomes small, which increases the *integration* region. That is, one is indeed likely to integrate traffic with similar QoS requirements when minimizing the bandwidth reservation.

One can look at the problem of whether to integrate or segregate from a different perspective. Suppose n_1 and n_2 are the numbers of Type1 and Type 2 connections carried by the system. Inequality (6) can then be written as:

$$\frac{n_1\sigma_1^2}{n_2\sigma_2^2} \geq \left[\frac{k_1^2 - k_2^2}{2k_1k_2}\right]^2.$$

In Fig. 5, we plot the *integration* and *segregation* regions with respect to n_1 and n_2 with aforementioned k_1, k_2 and $\frac{\sigma_1}{\sigma_2} = 0.6$. The figure clearly indicates that the ratio of n_1 and n_2, rather than their magnitudes, determines the decision.

These results show that accounting for the efficiency of multiplexing is essential in order to minimize the total bandwidth reservation. For cases with more than two traffic types, these observations still hold, but a simple criterion for the decision is unlikely because of the increased complexity. Nevertheless optimal VP formations of multiple traffic types could be determined numerically or by simulation. Indeed see [13] for an example in the case of on/off traffic. Finally we note that the ratio of bandwidth savings to total bandwidth requirement is proportional to $\frac{1}{\sqrt{N}}$ and thus becomes

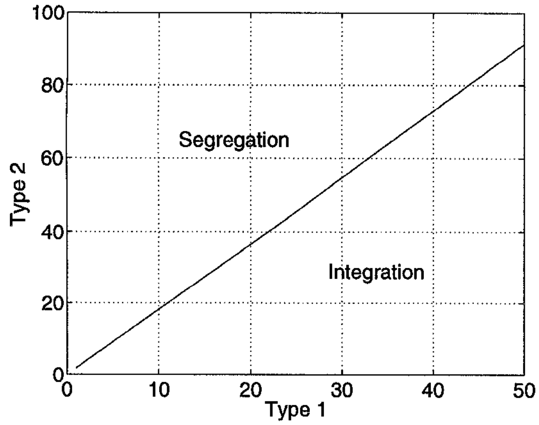

Figure 5: Integration and segregation regions.

less significant as system size increases, see [13].

4 Routing and Traffic Mixes

We have shown that it might be more efficient to carry multiple traffic types on the same VP. With efficient multiplexing, we minimize the total reserved bandwidth and in turn allow more connections to enter the system. However, using multiple VPs might be preferable (or necessary) in some circumstances. For example, we might choose to use multiple VPs through different links to increase reliability or due to capacity constraints. By doing so, we ensure that if a VP fails, the traffic can be quickly rerouted to other VPs and the performance degrades smoothly. Given such requirements, one needs to determine how to route heterogeneous traffic efficiently through multiple VPs.

4.1 Static Network Flow Problem

Suppose there are two VPs between an origin-destination pair. We first consider a simple static network flow problem for partitioning heterogeneous connections onto VPs, aiming to achieve good multiplexing. This problem is indeed an abstraction of routing permanent VCs in a VP network, where the goal is to minimize the total bandwidth reservation in order to leave more free capacity in the networks. For simplicity, we formulate the problem of routing two traffic (VC) types with Gaussian traffic models, but the results extend to a general set-up using the nonlinearity of admissible regions.

Consider two VPs with bandwidth c_1, c_2 and two types of Gaussian traffic streams as in §3. Assume there are n_1 Type 1 and n_2 Type 2 streams, and both have the same overflow probability requirement. We will consider two problems: first, whether this load is admissible, and second how to partition (or route[3]) these connections onto two VPs in order to minimize the total bandwidth requirement.

Suppose a fraction a of Type 1 and a fraction b of Type 2 traffic are sent to VP 1 and the remaining traffic is sent to VP 2, then the bandwidth requirements r_1, r_2 on each VP must satisfy the following inequalities:

$$\begin{aligned} c_1 &\geq r_1 = n_1 a m_1 + n_2 b m_2 + k\sqrt{V_1} && (7) \\ c_2 &\geq r_2 = n_2(1-a)m_1 + n_2(1-b)m_2 \\ &\qquad + k\sqrt{V_2}, && (8) \end{aligned}$$

where k is the QoS parameter, and $V_1 = n_1 a\sigma_1^2 + n_2 b\sigma_2^2$ and $V_2 = n_1(1-a)\sigma_1^2 + n_2(1-b)\sigma_2^2$ are the variances of the aggregate traffic on VP 1 and VP 2. The total capacity requirement is then given by $r_1 + r_2 = n_1 m_1 + n_1 m_2 + k(\sqrt{V_1} + \sqrt{V_2})$. An optimal partitioning policy is a pair of (a^*, b^*) such that (7) and (8) are satisfied, and $(\sqrt{V_1} + \sqrt{V_2})$ is minimized (or equivalently $r_1 + r_2$ is minimized).

Note that $V = V_1 + V_2 = n_1\sigma_1^2 + n_2\sigma_2^2$ is a constant representing the total *variance* of the aggregate traffic. We can represent V_1, V_2 as fractions of V, e.g., $V_1 = \alpha V$, $V_2 = (1-\alpha)V, \alpha \in [0,1]$. Since the contribution of m_1, m_2 to $r_1 + r_2$ is constant, the total bandwidth requirement is determined by the variance V_1, V_2 on each VP. Hence determining (a^*, b^*) is equivalent to picking V_1, V_2 or alternatively α such that $F(\alpha) = \sqrt{\alpha V} + \sqrt{(1-\alpha)V}$ is minimized. Given the shape of $F(\alpha)$ in Fig. 6, $F(\alpha)$ is minimized when $\alpha = 1$ or $\alpha = 0$, i.e., send all traffic to one VP or the other.

However, we must meet admissibility constraints (7) and (8), thus sending all traffic to the same VP might not be possible if c_1 or c_2 are not big enough. Nevertheless, based on Fig. 6, it is essential to keep α close to 1 or 0 in order to make $F(\alpha)$ small. In other words, a partitioning policy should distribute the total variance V in an *unbalanced* fashion so as to improve the efficiency of multiplexing.

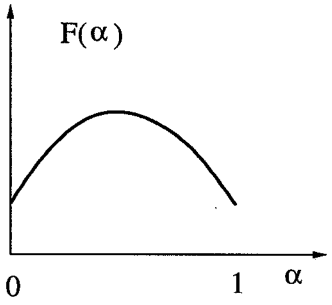

Figure 6: $F(\alpha)$.

[3] The terms routing and partitioning are used interchangeably in this discussion.

Joint admissibility. We have shown in §2 that a nonlinear admissible region can be obtained from the link's bandwidth and traffic statistics. In Fig. 7, we represent the joint admissible regions for two VPs: VP 1 on the first quadrant and VP 2 on the third quadrant. A point $S = (s_1, s_2)$ in the first quadrant represents a scenario where there are s_1 Type 1 and s_2 Type 2 streams in VP 1. Similarly a point $T = (-t_1, -t_2)$ in the third quadrant represents t_1 Type 1 and t_2 Type 2 streams in VP 2. A line segment $\overline{ST}$ connected by two points in the two regions represents $s_1 + t_1$ Type 1 streams and $s_2 + t_2$ Type 2 steams jointly admitted to the system.

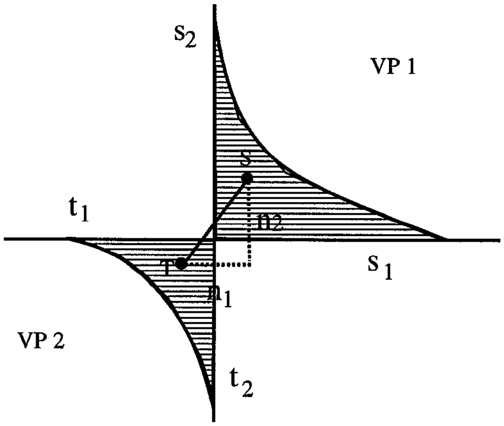

Figure 7: The joint admissible region.

Suppose $n_1 = Nf_1$ and $n_2 = Nf_2$, where $f_1 + f_2 = 1$. To determine whether (n_1, n_2) are jointly admissible, one needs to find points S and T such that $\frac{f_2}{f_1} = \frac{s_2+t_2}{s_1+t_1}$ and the length of $\overline{ST}$ is equal to $\sqrt{n_1^2 + n_2^2}$. In fact, every feasible $\overline{ST}$ with slope $\frac{f_2}{f_1}$ is associated with a partitioning policy for traffic with fractions f_1, f_2, and there exists one $\overline{ST}$ (may not be unique) having the largest length K^*. If $K^* \geq \sqrt{n_1^2 + n_2^2}$, then (n_1, n_2) are jointly admissible.

Lemma 4.1 *To accommodate a maximum number of connections with a given mix f_1, f_2, an allocation is made such that one of the two VPs carries homogeneous traffic, and the other VP carries mixed or homogeneous traffic.*

See [13] for proof.

Optimal routing. Suppose $\sqrt{n_1^2 + n_2^2} \leq K^*$, thus the heterogeneous traffic with combined number of connections (n_1, n_2) are jointly admissible. Now we consider the problem of partitioning them onto two VPs with a view on minimizing the total reserved bandwidth. We have shown that one needs to distribute variance V onto two VPs in an unbalanced fashion in order to increase multiplexing and minimize bandwidth reservation. Assume $c_1 > c_2$, then the objective is to determine the location of $\overline{ST}$ so as to make V_1 as large as possible. Since $V_1 = s_1\sigma_1^2 + s_2\sigma_2^2$, the point S of optimal $\overline{ST}$ will be on the admissible region boundary of VP 1 in order to make V_1 large.

Lemma 4.2 *To minimize the bandwidth reservation, one would pack the VP of larger bandwidth, and leave idle bandwidth, if any, on the other VP. One of the VPs will carry homogeneous traffic.*

See [13] for proof.

Lemma 4.2 suggests that the VP of larger bandwidth should be used as the *primary* VP, and traffic streams are packed into the *primary* VP, leaving the second VP partially occupied. The traffic fractions f_1, f_2 will affect how traffic streams are packed onto VPs in order to achieve good multiplexing.

4.2 Routing with Dynamic Call Arrivals

In the previous section we showed that a careful partitioning of heterogeneous connections can reduce the total bandwidth reservation because statistical multiplexing is affected by the traffic mix on each VP. Now we consider the VP routing problem in a dynamic set-up.

Suppose the call arrivals of the two traffic types are modeled by Poisson processes with rates $\lambda_1(\lambda_2)$ and each type of call has an exponential holding time with parameter $\mu_1(\mu_2)$ respectively. Instead of considering the admissibility or minimizing the bandwidth reservation, we consider the routing problem with the objective of minimizing the blocking probability. The results in Lemma 4.1 suggest that multiplexing is most efficient when each VP has unbalanced traffic mixes. Based on this observation, we propose a simple alternate routing algorithm which routes different traffic types to separate VPs. That is, each type of traffic is assigned a primary VP and a connection will be sent to its primary VP if it is available. If the primary VP is unavailable, the other VP is tried. The connection will be blocked if the second trial fails.

The proposed algorithms are compared with other algorithms under various traffic loads which are denoted $\rho_1 = \frac{\lambda_1}{\mu_1}$ and $\rho_2 = \frac{\lambda_2}{\mu_2}$ respectively. In the simulations, both traffic types are assumed to be on/off which have peak rates 1 and 0.5 with mean-to-peak ratios 10% and 90% respectively. The two VPs have an identical bandwidth 25, with admissible regions shown in Fig. 2. The routing algorithms we compare are shown in Fig. 8 and explained below.

A. Balanced-load scheme without re-trial. Connections of each type will be sent to each VP with equal probability. The loads sent to each VP are $\frac{1}{2}(\rho_1 + \rho_2)$.

B. Balanced-load scheme. The same as Algorithm A except if the selected VP is unavailable, the other VP is tried.

C. Aggregated-load scheme. A VP will be assigned as the primary VP for both traffic types. If the selected VP is unavailable, the other VP is tried.

D. The proposed algorithm. Different traffic types are assigned separate primary VPs, so the offered loads on each VP are ρ_1 and ρ_2. If the selected VP is unavailable, the other VP is tried.

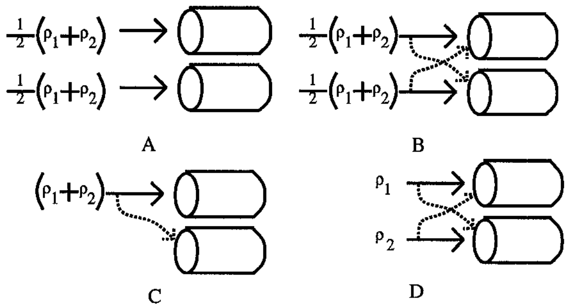
$\frac{1}{2}(\rho_1+\rho_2)$
$\frac{1}{2}(\rho_1+\rho_2)$
A
$\frac{1}{2}(\rho_1+\rho_2)$
$\frac{1}{2}(\rho_1+\rho_2)$
B
$(\rho_1+\rho_2)$
C
ρ_1
ρ_2
D

Figure 8: Four routing algorithms.

Simulation results. Based on the simulation results in Table 1, we found the Algorithm A has the worst overall blocking probability. This is because Algorithm A uses static load sharing without retrying the other path if the selected path is unavailable. Hence a connection may be blocked unnecessarily. The difference in blocking probabilities of Algorithm A and B strongly suggests that re-trial is worthwhile in order to improve the performance. Algorithm D has the smallest blocking probability under all traffic load combinations, which echos the observation found in the static network flow problem, that unbalanced traffic mixes improve the efficiency of statistical multiplexing. Intuitively an efficient multiplexing reduces the bandwidth reservation on each VP, which in turn reduces the chance of blocking. This is why the proposed algorithm has the best performance. Algorithm B and C have roughly the same blocking probabilities due to similar traffic mixes on the VPs. The traffic loads ratios on each VP of Algorithm B and C are roughly equal to the original offered loads ratio $\frac{\rho_1}{\rho_2}$, so the multiplexing is not as efficient as that in Algorithm D. From the results in Table 1, we can conclude that a careful flow partitioning will utilize the bandwidth of VP in a more efficient manner and further reduce the blocking. The proposed algorithm is robust under various traffic loads, and most importantly, it is simple to implement.

	A	B	C	D
ρ_1	*Blocking probabilities for Type 1*			
ρ_2	*Blocking probabilities for Type 2*			
40	0.0597	0.03099	0.03120	0.00221
40	0.0837	0.05027	0.04743	0.00222
45	0.0269	0.00791	0.00804	0.00044
30	0.0420	0.01477	0.01406	0.00048
30	0.0498	0.02280	0.02289	0.00148
45	0.0627	0.03149	0.03043	0.00149
50	0.0197	0.00462	0.00465	0.00069
25	0.0326	0.00913	0.00878	0.00095
25	0.0542	0.02796	0.02853	0.00289
50	0.0655	0.03405	0.03282	0.00290

Table 1: The comparison of blocking probabilities.

The choice of primary VP. We have shown that the proposed algorithm achieves an order of magnitude improvement in the blocking probability by sending different traffic types to separate VPs. The key remaining question is the choice of primary VP, i.e., which traffic type should be sent to which VP. We did not encounter this issue in the simulation results in Table 1 because the two VPs have the same bandwidth. Hence it will not make any difference how they are selected. However, this choice becomes important when the VPs have different capacities. We shall compare the blocking probabilities by simulating different choices of primary VPs under various traffic loads.

The simulation set-up is as follows. The two traffic types are those used previously in obtaining Table 1, and the bandwidth of VP 1 and VP 2 are 25 and 35 respectively. Based on intensive simulations, we found the "rule of thumb" is to assign the VP of large bandwidth as the primary VP of the traffic type which might result in the largest admissible number of homogeneous connections. In other words, the traffic with smaller "effective bandwidth" is sent to the larger VP. The heuristic is to carry a larger *number* of connections on the VP, so as to to exploit the "economies of scale." In particular, the reduction in blocking probabilities are more significant when the traffic with smaller "effective bandwidth" have higher offered load, e.g., see Table 2. Specifically when the traffic loads are $(100, 20)$, one can get almost an order of magnitude improvement in the blocking probability. Type 1 traffic has smaller effective bandwidth than Type 2 in the simulations, so Type 1 should use VP 2 as its primary VP. The results in Table 2 also suggest the same

Offered loads	Type 1 to VP 1	Type 2 to VP 1
(ρ_1,ρ_2)	Type 2 to VP 2	Type 1 to VP 2

(60,60)	p1=0.06607
	p2=0.08548
		p1=0.01894
		p2=0.02415

(70,50)	p1=0.05302
	p2=0.09209
		p1=0.00672
		p2=0.00859

(50,70)	p1=0.07594
	p2=0.07676
		p1=0.03626
		p2=0.04662

(80,40)	p1=0.03687
	p2=0.07563
		p1=0.00156
		p2=0.00186

(40,80)	p1=0.09131
	p2=0.08794
		p1=0.05813
		p2=0.06975

(100,20)	p1=0.00834
	p2=0.02108
		p1=0.00138
		p2=0.00296

Table 2: Blocking probabilities of different choices of primary VPs.

choice.

5 Conclusions

In this paper we have attempted to clarify problems related to resource allocation and routing in integrated services networks. The first natural question that arises is whether heterogeneous traffic with different QoS requirements should be segregated on distinct VPs. Based on a simple model our analysis shows that the answer is not straightforward. For this model a criterion for making such decisions is derived which depends on the traffic characteristics, traffic mixes, and QoS requirements. Similar behavior is likely to hold for more general setups, where the essential tradeoff is between achieving improved statistical multiplexing by aggregating but losing efficiency due to provisioning for the most stringent QoS.

Given a QoS requirement, such as cell loss at a link, the set of admissible numbers of connections of various types is very likely to have a nonlinear boundary. This reflects the role that the traffic mixes play in determining the effectiveness of statistical multiplexing of such traffic. The second question we considered addresses the impact of statistical multiplexing and relative traffic mixes in routing connections through the network. Indeed we show that in both a static and a dynamic routing model with heterogeneous traffic types, careful allocations or decisions can lead to a significant decrease in the required bandwidth or the blocking probability that connections will experience. The proposed, albeit simplistic routing algorithm, achieves robust performance over various traffic loads, as well as an order of magnitude improvement in blocking probability, over various other routing policies that do not explicitly deal with the heterogeneous character of the traffic load.

References

[1] P. Billingsley. *Probability and Measure.* John Willey and Sons, New York, 1986.

[2] S. Borst and D. Mitra. Asymptotically achievable performance in ATM networks. *To appear in Advanced Applied Probability.*

[3] A. Elwalid, D. Mitra, and R.H. Wentworth. A new approach for allocating buffers and bandwidth to heterogeneous, regulated traffic in an ATM node. *IEEE JSAC*, Vol. 13, No. 6:1115–1127, 1995.

[4] S. Gupta, K. W. Ross, and M. El Zarki. *Routing in Communications Networks*, chapter 2. Prentice Hall, ed. M. Steenstrup, 1995.

[5] J.Y. Hui. Resource allocation for broadband networks. *IEEE JSAC*, 6:1598–1608, 1988.

[6] J.Y. Hui. *Switching and Traffic Theory for Integrated Broadband Networks.* Kluwer Acad. Publ., Boston, 1990.

[7] F.P. Kelly. Routing and capacity allocation in networks with trunk reservation. *Mathematics of Operations Research*, Vol. 15, No. 4, 1990.

[8] F.P. Kelly. Effective bandwidths of multi-class queues. *Queueing Systems*, Vol. 9, No. 1:5–16, 1991.

[9] A. Papoulis. *Probability & Statistics.* Prentice-Hall, 1990.

[10] A. Shwartz and A. Weiss. *Large Deviations for Performance Analysis.* Chapman & Hall, London, UK, 1995.

[11] I. Sidhu and S. Jordan. Multiplexing gains in bit stream multiplexors. *IEEE Trans. Networking*, 3:785–797, 1995.

[12] R. Siebenhaar. Multiservice call blocking approximations for virtual path based ATM networks with CBR and VBR traffic. *IEEE Infocom 95*, pages 321–329, 1995.

[13] C.-F. Su. *Efficient traffic management based on deterministically constrained traffic flows.* PhD thesis, ECE Dept, Univ. of Texas at Austin, 1998.